Author's response

Mindreading underlies Metacognition

Peter Carruthers

Department of Philosophy, University of Maryland, College Park, MD 20742.
pcarruth@umd.edu
www.philosophy.umd.edu/Faculty/pcarruthers/

Abstract: This essay defends the view that human metacognition results from us turning our mindreading capacities upon ourselves, and that our access to our own propositional attitudes is through interpretation rather than introspection. Relevant evidence is considered, including that deriving from studies of childhood development and other animal species. Also discussed are data suggesting dissociations between metacognitive and mindreading capacities, especially in autism and schizophrenia.

R1. Introduction

The target article set out to consider four different accounts of the relationship between our mindreading and metacognitive abilities ("two independent mechanisms", "one mechanism, two modes of access", "metacognition is prior", and "mindreading is prior"). It argued in support of the fourth ("mindreading is prior") account, according to which metacognitive competence results from us turning our mindreading abilities upon ourselves. The target article considered a wide array of evidence bearing on the choice between the four accounts. This included evidence from childhood development, evidence from the role that metacognitive beliefs play in guiding human cognitive processes and behavior, evidence of confabulation in reports of one's own attitudes, alleged evidence of direct introspection of attitudes, comparative evidence of metacognitive competence in other species, evidence from autism, and evidence from schizophrenia. The commentaries take up an equally disparate set of topics. Some raise fundamental challenges that need to be confronted at some length whereas others (as might be expected) are based upon misunderstandings. Since there are few general themes linking the various commentaries together, Table 1 provides a guide to the structure of the present reply, together with an indication of which commentaries are responded to (in whole or in part) in each section (including the notes attached to that section).

Table 1

R#	Section	Commentaries
R2.	The nature of the mindreading faculty.	Buckner et al., Friedman & Petrashek, Lurz, Pereplyotchik.
R3.	The development of mindreading.	Anderson & Perlis, Buckner et al., Hernik et al., Lewis & Carpendale, Rochat.
R4.	The question of developmental priority.	Fernyhough, Mills & Danovich, Robbins.
R5.	What is introspection?	Baars, Murphy, Pereplyotchik, Petty & Briñol, Zinck et al.
R6.	Evidence for and against introspection.	Fiala & Nichols, Hurlburt, Zinck et al.
R7.	What is metacognition?	Anderson & Perlis, Couchman et al., Kornell et al., Proust.
R8.	Metacognition in animals?	Couchman et al., Kornell et al., Proust.
R9.	Dual processes and judgment.	Buckner et al., Frankish.
R10.	The evidence from autism.	Lombardo et al., Williams et al.
R11.	Neuroimaging evidence.	Lombardo et al.
R12.	The evidence from schizophrenia.	Robbins, Wiffen & David.
R13.	Some friendly suggestions.	Evans, Huebner & Dennett, Langland-Hassan.
R14.	Behaviorism bites back.	Catania, Lewis & Carpendale.
R15.	Conclusion.	—

R2. The nature of the mindreading faculty

In the target article I had hoped that my argument in support of a "mindreading is prior" account of self-knowledge was independent of any specific commitments concerning the character of the mindreading faculty itself, beyond rejection of a simulation-based "metacognition is prior" alternative. I still think that is partly correct. Certainly I can accept that the mindreading faculty isn't monolithic, but is actually a cluster of more specialized mechanisms working together in concert, somewhat as **Buckner et al.** suggest. Indeed, this is what I actually believe, following Baron-Cohen (1995) and Nichols & Stich (2003). But one powerful objection to the proposal that there is no such thing as introspection for attitudes, raised independently by **Friedman & Petrashek** and by **Lurz**, has made me realize that the argument cannot be free of all such commitments. I shall first outline the objection, before showing how an independently-motivated account of the architecture and mode of operation of the mindreading faculty can accommodate it.

The objection is that the mindreading system needs to have access to the agent's own beliefs in order to do its interpretative work; therefore self-attributing beliefs should be just as trivially easy as self-attributing experiences. **Friedman & Petrashek** claim, for example, that in order to form the metarepresentational belief that Bill believes that the first-aid box contains bandages, the mindreading system must access the attributor's own belief that first-aid boxes normally contain bandages. And they go on to stress that the mindreading system's default is to attribute the subject's own beliefs to other people, saying that this requires it to have access to those beliefs. Likewise, **Lurz** imagines a mindreader who observes a conspecific seeing some food. In order to draw any inferences from that fact, Lurz tells us, the mindreading system would have to access such beliefs as that the conspecific has recently eaten, or has shared food with others in the past. And again the moral is that the mindreading system needs to have access to the agent's own beliefs in order to do its work.

In light of these plausible claims, what might motivate one to deny that the mindreading system can access all of the agent's own beliefs? The answer is that the objectors forget about the frame problem. The idea that any single mental faculty might be conducting searches amongst all of a subject's beliefs is extremely problematic. Rather, there are likely to be a whole swarm of different decision-making systems that can conduct local searches of aspects of memory (Carruthers 2006). And a large part of the point of organizing cognition around a global workspace is that queries posted in that space can co-opt the resources of all the different consumer systems in parallel (Shanahan & Baars 2005). If the mindreading system is one of the consumer systems for global broadcast, as the target article assumes, then what we should predict is that it only has access to a limited database of domain-specific beliefs necessary to perform its computations.[1] But if this is so, then the challenge is to explain the datum that any one of one's beliefs can seemingly get appealed to in the course of mindreading.

To meet this challenge, I need to make two closely related distinctions. One is between System 1 mindreading (which is comparatively fast and done "online") and System 2 mindreading (which is slower and more reflective, and often involves supposition and simulation). This first distinction should need no defense. For reasoning about the minds of other people, like every other domain of reasoning that we know of, should admit of both System 1 and System 2 varieties. The other distinction is between verbally-mediated forms of mindreading (such as answering a question about what someone believes) from kinds of mindreading that don't involve access to linguistic representations. (We can be quite sure that the latter exist, since even

[1] And then to respond to **Lurz's** question why we should not believe that thoughts as well as perceptual states can be globally broadcast—raised also by **Pereplyotchik**—note that all of the evidence that we have of global broadcasting concerns perceptual or quasi-perceptual events. And note, too, that the best-established models of general-purpose working memory require the operation of one or another perceptual "slave system"—either the phonological loop or the visuo-spatial sketch-pad; see Baddeley (1990).

severely agrammatic aphasic people can retain their competence in non-verbal mindreading tasks. See Siegal & Varley 2002; Varley 1998.)

Consider, first, the fact that people will by default attribute their own beliefs to other people if asked. There is no reason to think that this requires the mindreading faculty to access those beliefs, any more than answering a question about one's *own* beliefs requires such access, as I argued in the target article (Section 2.1). Rather, the executive and language-production systems cooperate (and partly compete) with one another, searching the attributor's own memory and issuing the result in the form of a metarepresentational verbal report—"I think / he thinks that P"—where the form of the report can be copied from the form of the initial question. The mindreading system has the power to intervene in this process when it possesses a representation of the target's belief that differs from the subject's own, but it plays no part in the default attribution process itself. Consistent with this suggestion, Apperly et al. (2007) show that people are significantly slower when responding to a probe about a target's false belief than they are when responding to a reality-probe.

Now consider a reflective, System 2, instance of mindreading (whether verbal or non-verbal). A query about the target's thoughts, goals, or likely behavior is posted in the global workspace (either in the form of a verbal question, or in an image of oneself in the situation of the target, say). The entire suite of consumer systems gets to work, drawing inferences and reasoning in their normal way, accessing whichever of the subjects beliefs they normally would, and the results are posted back into the global workspace once more, where they are accessible to the mindreading faculty as input, perhaps issuing in a conclusion or a further query. Here the entire process, collectively, has access to all of the agent's beliefs; but the mindreading system has access only to whatever gets posted in the global workspace (in addition to its own domain-specific data-base, of course, which is accessible to it when processing).

Finally, consider a case of "on-line" unreflective System 1 mindreading, of the sort that might be engaged in by the infants in the false-belief studies conducted by Onishi & Baillargeon (2005), Southgate et al. (2007), or Surian et al. (2007). Perceptions of the main aspects of the unfolding events are attended to and globally broadcast, thereby being made available to the full range of conceptual systems including mindreading. These systems conceptualize and draw inferences from the input, with the former being fed back and broadcast as part of the perceptual state itself, and with the results of the latter being held briefly in the relevant domain-specific working memory system. (All System 1 reasoning systems will need to possess their own form of working memory, of course, to hold the results of previous computations while the next steps are undertaken. See Carruthers 2006.) Included in these broadcasts, then, will be the information that the target subject *sees* an object in one box rather than another, for example. And the working memory system that is internal to the mindreading faculty will contain such information as that

the target *expects* the object to be where it was last seen and is *ignorant* of the fact that it has been moved. When combined with a novel perceptual input (e.g. the target subject returns on the scene after a brief absence), these beliefs enable an expectation to be generated concerning the target's likely behavior.

Notice that on this account no beliefs need to be accessible to the mindreading system beyond those residing in its domain-specific data-base, with the exception of those that are made perceptually available to it, on the one hand, and those that are immediately-past products of its own operations, on the other. This is consistent with the fact that adults as well as children will fail to take account of the mental states of other people in their online reasoning once the relevant facts are no longer perceptually salient and sufficient time has elapsed for any record to have been expunged from the mindreading system's working memory. See Keysar et al. (2003) for a dramatic demonstration of this point.

R3. The development of mindreading

Just as I had hoped to make the argument of the target article largely independent of assumptions about the nature of the mindreading faculty, so I had hoped to minimize assumptions about the latter's development. (Of course I do need to assume that development does *not* begin with first-person awareness of our own attitudes, in the way that Goldman, 2006, suggests.) In particular, I tried to steer clear of the dispute between nativist or "core knowledge" approaches, on the one hand (e.g. Fodor 1992; Leslie et al. 2004) and constructivist or theorizing-theory accounts, on the other (e.g. Gopnik & Melzoff, 1997; Wellman 1990). Here, too, my efforts were partly, but by no means entirely, successful, as I shall now explain.

Both **Hernik et al.** and **Rochat** emphasize the crucial roles played by emotion and emotional engagement with others in the development of mindreading; and each appears to think that this claim conflicts with some aspect of the target article. But I see no problem with accepting these data. Both nativists and theorizing-theorists can believe in the developmental importance of emotional engagement, but will interpret the sources of that importance differently. Moreover, no evidence is provided that an understanding of one's own emotions precedes an understanding of the emotions of others in development (which would be problematic for a "mindreading is prior" account). On the contrary, Rochat writes: "From a developmental vantage point, affective reading and meta-affectivity are ontologically linked, representing two sides of the same coin." This is, of course, further grist for my mill.

Construed as a thesis about the architecture of the mature mind, the "mindreading is prior" account is independent of the debate between nativists and theorizing-theorists about the

development of the mindreading system. But **Buckner et al.** are correct in pointing out that one (though only one) of the *arguments* that I use in support of the "mindreading is prior" architecture depends upon some or other variety of nativist position (whether this be an innately given body of knowledge, or an innate module, or an innate domain-specific learning mechanism). For I claim that there is a good evolutionary explanation of the emergence of mindreading in highly social creatures such as ourselves, whereas there are no good evolutionary reasons for the emergence of introspection for attitudes (or else those reasons makes predictions that aren't borne out by the metacognitive data, either human or comparative). This is supposed to count in favor of a "mindreading is prior" account. And it plainly commits me to some or other version of nativism about the course of mindreading development.

Buckner et al. argue, in contrast, that metarepresentational mindreading may be a late exaptation of more primitive capacities, grounded in these together with our linguistic abilities and general-purpose concept-learning and theorizing skills. They think that the only true adaptations in the social-cognitive domain are a swarm of first-order, non-metarepresentational, mechanisms for face recognition, eye-tracking, automated imitation via the mirror neuron system, and so forth. But there are two main problems with this view. One is the rapidly expanding body of evidence of *very* early metarepresentational competence in infants, embracing false-belief understanding inter alia (Bosco et al. 2006; Onishi & Baillargeon 2005; Onishi et al. 2007; Song & Baillargeon forthcoming; Song et al. forthcoming; Southgate et al. 2007; Surian et al. 2007). And not all of these studies, it should be stressed, use looking time as a measure of expectation violation. On the contrary, Southgate et al. (2007) use anticipatory looking as their dependent measure, which is much less ambiguous.

The other major problem with **Buckner et al.'s** suggestion is that mindreading is required in order to learn a language in the first place. I don't deny that syntax may be innate, or acquired through the offices of a dedicated domain-specific learning mechanism. But learning the lexicon requires children to figure out the referential intentions of the speakers around them (Bloom 2002). And this plainly requires metarepresentation. Moreover (and just as this account predicts), we have ample evidence that infants can attribute goals and intentions to others in the first year of life, significantly before they can attribute beliefs and misleading appearances (Csibra et al. 2003; Johnson 2000; Luo & Baillargeon 2005; Woodward 1998).

Buckner et al. write admiringly of the work of Gallagher (2001, 2004) in this connection, as do **Anderson & Perlis** and **Lewis & Carpendale**. But Gallagher's work is subject to both of the objections outlined above. Moreover, he goes awry in his critique of the opposing approach, which he refers to with the generic "theory-theory" (intended to cover both nativist and theorizing-theory varieties). In particular, it is simply false that theory-theorists must (or do) assume that mentalizing usually involves the adoption of a third-person, detached and

observational, perspective on other people. On the contrary, theory-theorists have always emphasized that the primary use of mindreading is in *interaction* with others (which Gallagher calls "second-personal"). That, after all, is what "Machiavellian intelligence" is all about. And the fact that our apprehension of the meaning of other people's behavior is often phenomenologically immediate does not, of course, show that it isn't underpinned by theory-driven computations of underlying mental states. Indeed, there is simply no other way of explaining our competence in this domain. Appealing just to sensory-motor skills (as Gallagher does) is plainly inadequate to account for the *flexibility* of the ways in which adults and infants can interact with others. Indeed, in order to interact flexibly with *any* complex system (be it physical or human) you need a good enough understanding of how it works.

R4. The question of developmental priority

The target article expressed skepticism about the capacity of developmental data to discriminate between a "mindreading is prior" account and its three main competitors (Section 4). **Mills & Danovich** disagree. They cite a number of forms of evidence suggesting that mindreading skills of various sorts emerge in development prior to metacognition, which they say supports a "mindreading is prior" account. Since one of my main grounds for skepticism concerned arguments for the priority of mindreading that are premised on the *parallel* emergence of mindreading and metacognition in development (which fails to discriminate between the "mindreading is prior" view and the "one mechanism, two modes of access" account), I am happy to agree. But let me sound the following cautionary note. Until we have a good understanding of the reasons for the two-year developmental lag between children's capacities to pass non-verbal and verbal versions of mindreading tasks, arguments that rely upon the latter need to be treated with some caution. For it may be that the "self" and "other" versions of a verbal task differ along whatever turns out to be the relevant parameter. Put differently: you can't control for confounding factors that you don't yet know about.

In response to **Mills & Danovich** I should also stress that while a finding that mindreading competence is developmentally prior to metacognition would support a "mindreading is prior" account (because it would be inconsistent with the other three alternatives), this isn't actually a *prediction* of the account. For the latter claims only that it is *the same system* that underlies our mindreading capacity that gets turned upon ourselves to issue in metacognition. It doesn't claim that the first occurs in development before the latter. (In this respect the label "mindreading is prior" may be somewhat misleading. I intend it only to refer to a *functional* and/or *evolutionary* priority.)

Fernyhough would plainly disagree with the point made in the previous paragraph. He gives

reasons for thinking that it may take time for aspects of children's inner lives to develop. In particular, the transformation of private speech ("talking to oneself") into *inner* ("silent") speech may not be complete until middle childhood; and capacities to build and sustain visual images may likewise be slow to develop. Since the target article claims that these are amongst the data that the mindreading system uses when attributing propositional attitudes to oneself, Fernyhough says that the "mindreading is prior" account must therefore predict that metacognition should lag significantly behind mindreading in development. But there is no such implication. All that follows is that there will be many more moments in the daily lives of children at which they will be unwilling to attribute occurrent thoughts to themselves than is true of the daily lives of adults, because the conscious mental events that might underlie such self-attributions simply aren't present. Nothing follows about children's *competence* to self-attribute attitudes. Nor does it follow that children will be weaker at attributing attitudes to themselves than they are at attributing attitudes to others, provided that the tasks are suitably matched.

Robbins claims that I have overlooked crucial conflicting evidence, which demonstrates that metacognition is prior to mindreading in development. He cites a study by Wimmer et al. (1998) which seems to show that young children have awareness of their own knowledge before they have awareness of the knowledge of other people. But the study in question admits of an alternative explanation. In the "self" condition, the children are allowed to look, or not look, into a box, and are then asked whether they know what is in the box; whereas in the "other" condition they observe a subject either looking, or not looking, into the box before being asked whether the subject knows what is in the box. Answering the question in the "other" condition requires the children to reason appropriately from the generalization that seeing leads to knowing (or somesuch). But answering the question in the "self" condition requires no such thing. The children can answer simply by accessing, or by failing to access, their knowledge of what is in the box. They can substitute a first-order question in place of the second-order question asked—namely, "What *is* in the box?"—and answer "Yes", that they do know what is in the box, if an answer comes to mind, otherwise answering "No".

R5. What is introspection?

Baars thinks that the target article is committed to denying that nonhuman animals and young infants feel pain. This is because these subjects are incapable of mindreading, and because the target article denies the existence of introspection. But there are two distinct misunderstandings at work here.

One is based upon an unfortunate ambiguity in the use of the term "introspection". In one sense, introspection is any form of looking within *the body*. In this sense perceptions of pain or of one's

own beating heart count as introspections. In another sense, introspection is a form of looking within *the mind*. In this sense the outputs of an introspective process are always metarepresentational, involving representations of one's mental states as such. And in this sense perceptions of pain or of heart-beat are definitely *not* introspections, since they issue in first-order representations of properties of the body. It should be stressed that it is only this latter, metarepresentational, sense of "introspection" that is at stake in the target article. Hence even if I denied the existence of introspection in this sense altogether, there is no reason why this should commit me to denying that animals feel pain, or fear, or hunger, or thirst. For what is in question in these cases is only introspection in the first "within the body" sense.

Baars' second misunderstanding lies in believing that the target article denies the existence of introspection (in the metacognitive sense) for all categories of mental state. For he thinks that the view will have difficulty in accounting for the reliability of metacognitive self-report in psychophysics. But I specifically allow (indeed, I insist) that globally broadcast perceptual and quasi-perceptual states can be introspected, because they are available as input to the mindreading faculty. Self-attribution of such states should therefore be trivial for anyone who possesses the requisite concepts, which can then be applied to the input-states on a recognitional (non-interpretative) basis.

Pereplyotchik, too, misunderstands the sense of "introspection" that is at issue. For he thinks that it will be sufficient to demonstrate that there is no such thing as introspection for perceptual states if it can be shown that the mindreading system relies upon a tacit theory in self-ascribing such states. This is a mistake. That a process is introspective is *not* supposed to be inconsistent with it involving computations or inferences of various sorts (provided that they are unconscious ones), so long as the inferences rely only on information of a general kind, and don't access information about the agent's circumstances, behavior, or earlier mental states. For remember, what is at stake is whether our access to our own minds is different *in kind* from our access to the minds of other people. And the latter always involves just such inferences. This was also the reason why I defined introspection *negatively* for the purposes of the target article. For I wanted to leave open "inner sense" accounts as well as "application of a tacit theory" views of introspection, according to each of which the attribution of mental states to oneself is inferential (but still quite different from the attribution of mental states to other people).

Zinck et al. mistake the nature of the intended contrast between a "one system, two modes of access" account and a "mindreading is prior" view. They insist that when metarepresenting our own mental states the mindreading system has access to a richer array of data, such as visceral and somaesthetic sensations, and that this therefore supports a "one system, two modes of access" account. But I, too, claim that the mindreading system can utilize data when attributing states to the self that aren't available when attributing the same states to others, and maintain that

this is consistent with a "mindreading is prior" view. As I intend the distinction, the difference between the two forms of account is *not* whether there are different *data* available to the mindreading system when attributing mental states to oneself or to another. Rather, the difference concerns whether the mindreading system employs two different *informational channels* in the two cases. The distinction is intended to be an architectural one. Since the mindreading system utilizes the very same mechanism of "global broadcast" of attended outputs of perceptual systems, whether attributing mental states to oneself or to another, this means that there are *not* two different modes of access to mental states, even though the perceptual and quasi-perceptual states that are utilized in the two cases are often different. To provide evidence supporting a "one system, two modes of access" account, Zinck et al. would need to show that we can self-attribute propositional attitude states independently of any sensory or imagistic information accessible via global broadcast. But they themselves seem to doubt whether any such evidence exists.

Murphy denies that the mindreading system is always implicated in our knowledge of our own attitudes, while agreeing with me that there are no special mechanisms that enable us to detect and describe those attitudes. Rather, he thinks that we can do whatever we would normally do to determine a truth about the world and can then use the result of that same process to self-ascribe the resulting belief. This might well work as an account of how we express our beliefs in speech. Indeed, so construed, it is an account that I endorsed in the target article (Section 2.1). The language production system can take as input the result of a belief-forming process, or the result of a search of memory, and can formulate that input-state into a belief report. We can imagine this happening via a two-step process: the language system accesses a belief with the content *P* and draws on lexical and syntactic resources to express this in a sentence, "P", before elaborating it and articulating the result in the form, "I believe that P." But I deny that such an account can succeed as an account of meta*cognition*, or as an account of how we form *beliefs* about our own beliefs.

Murphy is confronted with the following dilemma. Suppose, first, that the assertion, "I believe that P" is an encoding into language of a previously existing metacognitive belief (the belief, namely, that I believe that P). Then the challenge is to explain how this belief is arrived at without either implicating the mindreading faculty or appealing to any special introspective channel. But there would seem to be just two possible ways for whatever process issues in such a belief do its work (in a reliable enough way). One would be for it to have access to the output of the process that issues in the belief or memory that P (which would then surely involve some sort of introspective channel of access to the latter). The other would be for the metacognitive belief-forming process to involve interpretation and inference from other events, such as a prior tokening of the assertion, "P", or the occurrence of a memory-image caused by the belief in question (which would surely then implicate the mindreading faculty, or else some system with

many of the same powers as the mindreading faculty).

So **Murphy** must intend (much more plausibly) that the assertion, "I believe that P" can be generated directly from the belief that P, without subjects first needing to form the metacognitive belief that they believe that P. As described above, the language system (working in concert with executive systems, no doubt) can access the belief that P but then formulate this into the sentence, "I believe that P", rather than the first-order sentence, "P". But this assertion isn't *itself* a metacognitive belief (nor, by hypothesis, does it involve one). Rather, it is a linguistic action (albeit one with a metarepresentational content). The system that issues in the metacognitive belief that I believe that P must take this assertion as input and deliver the metacognitive belief as output. But in order to do this, it would have to engage in interpretation, just as when hearing a similar assertion made by another person. Since the assertion could be a lie, or meant ironically, or meant as a joke, it is hard to see how the necessary interpreting could be done except by the mindreading faculty (or else some system with many of the same powers as the mindreading faculty). But this is now the view endorsed by the target article: in such cases I come to know what I believe by hearing and interpreting what I say (whether overtly or in inner speech). Murphy has failed to present us with a genuine alternative.

Petty & Briñol agree with the target article that self-attributions of attitudes always involve interpretation. But they insist that interpretation is a matter of *degree*, and that sometimes interpretation can be so minimal as to be almost indistinguishable from introspection. I agree with the former point but not with the latter. Of course it is true, as Petty & Briñol point out, that there is a big difference between interpretations of oneself that rely only on publicly available information (such as one's own behavior and circumstances) and interpretations that rely only on subjectively-accessible mental events (such as one's own somatic feelings and/or one's own "inner speech"). But the main point at issue in the target article *is* a dichotomous, architectural, one. It concerns the existence (or not) of a distinct informational channel to our own attitudes, different from the sensory channels that are available to the mindreading system for use in interpreting other people. There either is such a channel or there isn't. (The target article claims the latter, and Petty & Briñol appear to agree.) Moreover, even minimal-interpretation cases are much less similar to introspection than Petty & Briñol seem to think. Consider their example of someone who says to himself, "It is good", when tasting some ice-cream, and thereby interprets himself as liking ice-cream. The mindreading faculty, functioning together with the language comprehension system, has to fix on the object of evaluation ("*What* is good?"), interpret the evaluative predicate ("In what sense is it good?"), and determine what sort of speech act is being expressed (whether literal, suppositional, ironic, or whatever). No doubt the answers can, in context, be settled quite easily. But they are exactly *the same* answers that would need to be provided when interpreting the speech of another person. And no one should think that the latter is at all similar in its nature to introspection.

R6. Evidence for and against introspection

Fiala & Nichols challenge the claim made in the target article that confabulators often have the impression that they are introspecting rather than self-interpreting (Section 3.1), which is a crucial component of the argument against introspection for attitudes. They first point out that no one has ever *asked* a split-brain subject whether or not he thinks he is introspecting. But this would be a bad question to ask, for a number of reasons. One is that "introspection" is a term of art, and requiring people to make judgments involving an unfamiliar term is unlikely to be a reliable way of finding out what they believe. Another is that the direct-question method is a poor way of accessing people's tacit beliefs in general (Scholl 2007). I doubt that many people have explicit, verbalizable, beliefs about the nature of their access to their own mental states—with the possible exception of those who have taken an introductory course in philosophy. Rather, the way in which people think and reason about their own mental states just *assumes* that the latter are transparently accessible to them. But if *asked* about that access, who knows what they might say? For they will almost certainly find the question to be confusing, and they might revert to bits and pieces of knowledge acquired about Freud, or about cognitive science, or whatever, when trying to say something sensible by way of answer.

So what is really in question is whether it seems to split-brain subjects that they are formulating beliefs about their own mental states and processes in whatever way they normally would—in a way that doesn't seem to them to involve self-interpretation—not whether they have explicit beliefs about the process in question. This is hard to assess directly. But those who work with such people say that their own sense of themselves following the split-brain operation seems to be unchanged (Gazzaniga 1995). And even reminders of their split-brain status that are made immediately prior to testing—and that are given, moreover, to those who have a good theoretical understanding of the effects of the operation—have no effect (Gazzaniga, email communication 11/8/06). The subject goes right on confabulating. This isn't what one would predict if subjects were, at any level, aware of interpreting themselves, since one would expect that a reminder of their split-brain status should enrich their hypothesis pool. But it doesn't.

Fiala & Nichols also point out that there are many examples from the confabulation literature where subjects express their metacognitive thoughts with low confidence, suggesting that they are not only interpreting themselves, but are at some level aware that they are doing so. The point is entirely correct. But it doesn't have the consequences destructive of my argument that Fiala & Nichols allege. This is because there are also a great many instances in which subjects express their metacognitive beliefs unhesitatingly and with *high* confidence. And these are all that I require to make my case. Indeed, the self-interpretative model of attitude self-awareness

predicts that there should be cases of both sorts. For only if an interpretation can be arrived at smoothly and unhesitatingly will subjects have the impression that they are introspecting. In more problematic cases such as those that Fiala & Nichols cite, or such as especially bizarre actions performed following hypnosis, it will be more difficult for the mindreading system to generate an interpretation (just as it would be difficult to interpret such behavior observed in another). And as soon as subjects become aware of themselves as interpreting, they are likely to express any belief that they formulate with some caution.

Note that exactly the same distinction can be made with respect to other-person mindreading. In many cases the interpretation process is swift and unconscious, and the resulting phenomenology is that we just seem to *see* someone's behavior as informed by certain beliefs and goals. (Here I am in full agreement with **Zinck et al.**) But in other cases an interpretation is harder to come by, and we become aware that we are trying to interpret. (See also the discussion of System 1 versus System 2 mindreading in R2.)

In the target article I assumed that one of the biggest challenges to a "mindreading is prior" account derives from the "descriptive experience sampling" studies conducted over the years by Hurlburt and colleagues (Hurlburt 1990, 1993; Hurlburt & Akhter 2008; Hurlburt & Heavey 2006), specifically the finding that subjects will sometimes report engaging in "unsymbolized thinking" at the time of the beep. I took this to be evidence that subjects are capable of introspecting their propositional attitudes, and tried to respond. However, **Hurlburt** now replies that I have misinterpreted his position. Unsymbolized thoughts are merely thoughts that don't have any semantically-relevant images, words, or other sensations as the "primary theme or focus" of the subject's attention at the time of the beep. Hurlburt concedes that such experiences are generally present in the periphery of attention, providing a basis for self-interpretation. Moreover, he argues that the ways in which subjects respond when probed about these episodes actually speaks *in favor* of a "mindreading is prior" position on our awareness of our own attitudes. This additional support from an unexpected quarter is, of course, most welcome.

R7. What is metacognition?

A number of commentators accuse me of using the term "metacognition" in a non-standard sense (**Anderson & Perlis**, **Couchman et al.**, **Proust**).[2] These commentators allege that the normal usage in cognitive science is that metacognition is involved in any process that has a controlling influence on the way that another cognitive process unfolds. On this account, it is left open

[2] While **Kornell et al.** make a similar claim, in their case it is based on a misreading of my own view. So far as I can tell, they mean by "metacognition" precisely what I do.

whether or not metacognition need involve metarepresentations of the events within the cognitive process that gets controlled.

I am happy to allow that some authors might use the term in this (hereafter "control") sense. But I deny that it is a common—let alone a standard—usage. In general in the metacognition literature in psychology, metacognition is defined in terms of thought about our own thoughts. Indeed, Proust herself provides the standard definition (2007, p271): "This is the domain of metacognition: thinking about one's own thinking." (See also Dunlosky & Metcalfe 2009; Flavell 1979; Koriat 2007.) And it is then a matter of substantive *investigation* whether or not, and to what extent, metacognition has a controlling function. (See especially Koriat et al. 2006.) This wouldn't even make sense if metacognition were *defined* in terms of control.

It is important to emphasize that the control and metarepresentational senses of "metacognition" are two-way independent of one another. There are certainly many instances in which one cognitive process exercises a causal influence on another without the former involving any metarepresentations of any aspect of the latter. (See Section 5.1 of the target article for some examples.) And in connection with any metarepresentational form of metacognition, it will always be an open question whether or not it has a causal influence upon the cognitive state or process represented. While these points are very well understood by most researchers, some are apt to think that they can move freely from talk of metacognition in the control sense to metacognition in the metarepresentational sense. This is especially true of some of those who work in the field of animal metacognition. Some, I think, are quite clear-headed that they are seeking forms of metacognitive control for which the *best available explanation* will be the occurrence of a metarepresentational process. (See especially Metcalfe 2008; Son & Kornell 2005.) But some seem unaware that any additional argumentation is needed to get from metacognition in the control sense to metacognition in the metarepresentational sense. This is especially true of the commentary by **Couchman et al**., as well as the articles by members of their team cited therein, which will be discussed in R8.

Proust raises a more substantive challenge to the assumptions of the target article. She suggests that the latter overlooks the possibility of *nonconceptual* forms of metacognition (in the metarepresentational sense of the latter term). Specifically, she suggests that epistemic feelings like surprise and confidence should be seen as nonconceptual representations of the underlying mental states (such as violated expectations or high degrees of belief). Hence any person or animal that can use such feelings as a *cue* to guide further behavior (such as looking more closely at the target event) can be said to be acting as a result of a metacognitive process. This is an interesting idea, which deserves examination. It will require us to delve a bit into competing theories of the nature of intentional, or representational, content.

Let us assume (with **Proust**) that epistemic feelings like surprise and confidence are distinctive forms of somatosensory experience that are caused by an underlying cognitive state or process, but without involving any conceptualization of that state or process as such. So an animal that feels surprise has an expectation (a belief) that is violated by what it is currently perceiving, which in turn causes a suite of bodily reactions of which the animal is aware (heightened alertness, widening of the eyes, automatic orienting towards the stimulus, and so on), but without the animal necessarily knowing *that* it has an expectation that has been violated. Since the epistemic feeling is reliably caused by a cognitive state or event, it thereby carries information about it. And then on any purely informational account of representational content (e.g. Fodor 1990), the feeling can count as a nonconceptual representation of the representational state or event in question (that is, it counts as a metarepresentation). One problem with this proposal, however, is that it makes metarepresentations come too cheap. For almost all mental states, processes, and behaviors will carry information about the existence of some other mental state or process, thereby becoming nonconceptual metarepresentations of the latter, on the proposed account. Thus inferential processes will characteristically carry information about (and hence metarepresent) the presence of beliefs, decision-making processes will carry information about the presence of beliefs and desires, and so forth.

Moreover, few researchers in cognitive science actually rely upon an informational account of representation in their own work. Most adopt some or other variety of inferential or conceptual role semantics (e.g. Block 1986), according to which what a symbol represents depends (at least partly) upon the use that the rest of the cognitive system is apt to make of that symbol. This is probably wise, since purely informational accounts of intentional content face notorious difficulties (one of which will be mentioned below; see Botterill & Carruthers 1999 for discussion). And then the question for us becomes: does the animal *make use of* the epistemic feeling in question in such a way that the feeling is thereby constituted as a nonconceptual representation of a cognitive state?

Consider, first, paradigmatic cases of nonconceptual representation, such as a perceptual representation of a colored surface or of the detailed shape of an object. In virtue of what does the perceptual state represent the colored surface rather than, for instance, a particular pattern of activity on the retina or in the optic nerve (since it equally carries information about both)? A natural answer is that the animal itself *treats* that representation as a representation of color—it thinks and acts in the sort of way that would be appropriate if it *were* a representation of the color of a surface. For example, perceiving the red surface of a fruit, and believing that red fruits of that type are ripe, the animal might grasp and eat it. Likewise a perceptual representation of the detailed shape of an object will be used to guide the animal's choice of grip size and hand orientation when it reaches out for it. It seems that a nonconceptual representation of some property of the world represents what it does partly in virtue of its role in guiding thought and

action that is focused on that aspect of the world.

Consider, now, epistemic feelings, such as the feeling of low confidence that an animal might experience when faced with an especially difficult judgment or discrimination. This is a feeling that involves an aversive state of anxiety, caused by the animal's low degree of belief. Should it be considered a nonconceptual representation of a cognitive state (one of low degrees of belief or of conflicts of belief), as **Proust** suggests? To answer, we need to look at how it is used by the animal. One thing that the animal might do in consequence is opt for a high-confidence, low anxiety, option instead. But this is an action that is targeted *on the world* rather than on the animal's own beliefs. It should lead us to say that the feeling of uncertainty is a representation of the riskiness of certain worldly options or events, rather than a representation of the animal's own low degree of belief. For the animal doesn't act in a way that is directed at its own beliefs; rather it acts on the world. Likewise for an animal that is led by its feeling of uncertainty to engage in information-seeking behavior such as examining the object more closely, walking around it to look at it from the other side, sniffing it, pressing a "hint" key of the sort employed by Kornell et al. (2007), and so on: these are behaviors that are aimed at answering a first-order question about the object—"Is it edible?", " Is it safe?", "What comes next?", etc.—rather than being aimed at changing the animal's own degrees of belief. It seems reasonable to conclude, therefore, that epistemic feelings should not be regarded as possessing metarepresentational nonconceptual content.[3]

Moreover, there is no reason to think that epistemic feelings are a first evolutionary step on the road to metarepresentation. This is because metarepresentation requires the development of concept-wielding consumer systems for the bodily cues in question, which contain implicit or explicit theories of the nature and causal roles of the underlying mental states. (Note that even a simulation theorist like Goldman (2006) needs to postulate an innately structured set of representations in a language of thought linked up to the different mental state kinds.) It should be stressed that the bodily feelings in question—that are distinctive of surprise, or the anxiety that attends uncertainty, for examples—are just that: bodily feelings. By themselves they give no clues as to the nature of the mental states that cause them (a violated expectation, in the case of surprise, and low or conflicting degrees of belief, in the case of uncertainty). How would an animal that as yet had no conception of those types of mental state be expected to acquire one? Certainly not via individual learning, surely. And if via evolution, then it is far from clear where the pressure to develop such theories is to come from. Not from the benefits of metacognition in the control sense, presumably, since by hypothesis the animals in question already have that (see

[3] Note that the behavior-guidance account of representational content proposed by **Anderson & Perlis** will also have exactly this consequence, since epistemic feelings guide action targeted on the world rather than on the animal's own mental states.

R8 below). Hence the pressure is presumably social, in which case what develops will be a mindreading system (albeit one that is capable of taking bodily cues as input).

R8. Animal metacognition?

There are two distinct ways in which an animal might behave in metacognitive fashion (in the control sense) without engaging in metacognition (in the metarepresentational sense). First, it might utilize degrees of belief and desire (without metarepresenting them as such, of course), combined with one or two simple first-order (non-metarepresentational) mechanisms and/or acquired first-order beliefs. This is the explanatory strategy followed in Carruthers (2008a) and described briefly in the target article. But second, as **Proust** explains, an animal might utilize its own bodily changes and reactions (including feelings that are distinctive of surprise, uncertainty, and familiarity, for examples) as *cues*. Thus an animal might be innately predisposed, or might have learned, that the bodily feeling distinctive of uncertainty is a cue to engage in actions that obtain improved information (e.g. by moving its head from side to side for a better view), or to engage in actions that have the effect of "opting out" of the current situation and entering a new one. (Actually, this might be one way of *implementing* the "gate-keeping" mechanism postulated in Carruthers 2008a, which takes roughly equivalently motivated but incompatible plans of action or inaction as input, and which motivates a search for additional information or alternative behavior.) Note that on this second alternative, the animal doesn't have to *recognize that* it is surprised or uncertain; indeed it needn't even possess the concepts of surprise or uncertainty. Rather, it just has to be capable of recognizing a distinctive bodily gestalt or feeling, and initiating an innately prepared or learned response. (Nor, as R7 argues in reply to Proust, does the feeling itself possess a nonconceptual metacognitive content.) Such an animal might display any of the metacognitive control-behaviors currently investigated by comparative psychologists, as I shall show briefly below. But it might be wholly incapable of metacognition in the metarepresentational sense.

Carruthers (2008a) describes exhaustively and in detail how all of the data cited in the commentary by **Kornell et al.** are explicable in non-metarepresentational terms, utilizing *degrees* of attitude strength. Kornell et al. make no attempt in their commentary to respond to those arguments, or to demonstrate why their own metarepresentational interpretation of their data is superior. It wouldn't be fruitful just to repeat the arguments of my 2008a here. Rather, I shall show briefly how an animal that can treat its own epistemic feelings as a cue might behave in the ways that Kornell et al. describe without being capable of metarepresentation. Thereafter I shall show how **Couchman et al.** chronically conflate the control and metarepresentational senses of metacognition in their commentary and other work.

Consider what is perhaps the most challenging set of data presented by **Kornell et al.**, citing Kornell et al. (2007). Rhesus monkeys were initially trained in a gambling task that required them to first make a difficult perceptual discrimination before choosing between a "high risk" and a "low risk" symbol. Pressing the former would issue in a large reward if the initial discrimination was correct but a large loss if the discrimination was incorrect; pressing the latter would issue in a guaranteed small reward. The monkeys were then trained on a quite different—memory—task (initially without the gambling option). Following training, when the "high risk" and "low risk" symbols were re-introduced, the animals immediately made adaptive use of them. Kornell et al. take this to show that the monkeys had learned a general metacognitive strategy for solving the tasks. Carruthers (2008a) shows how these data can be better explained in terms of degrees of belief combined with a capacity for abstract rule-formation. Here let me sketch a further alternative: that the animals might have learned to use their own feelings of uncertainty as a cue.

We can presume that monkeys are capable of both being, and *feeling*, uncertain, even if they are incapable of metarepresentation of any sort. The monkeys in the first phase of the above experiment could then have learned to treat their own feeling of uncertainty when making an initial discrimination as a cue to press the "low risk" symbol thereafter. They would therefore have acquired, and learned to act upon, a rule of the form, "When *that* bodily feeling/gestalt is present, press the 'low risk' symbol when it arrives." (Note that there is nothing metarepresentational contained here. The feeling in question is a state of the body, not of the mind. See R5 and R7.) When the monkeys then entered the second phase of the experiment they would, of course, sometimes feel uncertain, but this time whenever presented with a difficult *memory* task. The introduction of the gambling option might then have activated, and led them to act upon, the exact same rule.

I now turn to consider **Couchman et al.** It is plain that at the outset of their commentary they actually use "first-order" to mean "behaviorist", and that by "metacognitive" they mean any process that is genuinely *cognitive*, with the animal taking decisions in light of its beliefs. For they describe Smith et al. (2006) as supporting a "metacognitive" account. In those experiments both feedback and rewards were deferred until the animal had completed a block of trials, thus preventing the creation of stimulus-response pairings that might otherwise explain the animals' adaptive use of the uncertainty response. Couchman et al. write, "It was clear in that study that monkeys' uncertainty-response strategies were adjudicated cognitively and decisionally, not using first-order cues." I agree (at least, if by "first-order cues" one means, "stimulus-response pairings"). But the training would have given the animals ample opportunity to acquire a set of non-metarepresentational beliefs about the contingencies of the experiment. By the time that they entered the test phase, they would know that pressing the "dense" key if the stimulus was dense would thereafter issue in a reward, whereas pressing the "dense" key if the stimulus was sparse

would thereafter issue in a penalty, and that pressing the "uncertain" key would issue in neither a reward nor a penalty. These beliefs, combined with *degrees* of belief that a given stimulus is dense, or sparse, can then explain the data in an entirely non-metarepresentational way, as Carruthers (2008a) demonstrates.

Couchman et al. point out, quite correctly, that the non-metarepresentational explanation adverted to above requires the postulation of what Carruthers (2008a) calls a "gate-keeping mechanism" (which might be absent in capuchins and pigeons, note, thus accommodating the findings of Beran et al., 2008, and Inman & Shettleworth, 1999, that neither species makes adaptive use of an uncertainty response). This is a mechanism that is sensitive to the presence of beliefs or motivations for action of roughly equal strength, issuing in a search for additional information or alternative strategies when receiving such states as input. Couchman et al. object that this commits me to a metacognitive explanation of the data, and they write, "It [the gatekeeper mechanism] meets the definition of a second-order, controlled cognitive process." Since it is plain that the mechanism in question need not involve any metarepresentations for it to operate as envisaged, Couchman et al. must here be using "metacognitive" in the control rather than the metarepresentational sense.

So far there isn't any substantive disagreement between **Couchman et al.** and myself, just "crossed wires" resulting from differences in the use of the term, "metacognitive". But they go on to conclude their commentary by claiming victory for a "metacognition is prior" account over my own "mindreading is prior" model, despite the fact that the two are perfectly consistent with one another if the former is taken in their control sense and the latter is understood in my metarepresentational sense. They also offer an account of the origins of mindreading that is blatantly and explicitly Cartesian, presupposing that we have prior awareness and understanding of our own mental states *as such* (i.e. presupposing the prior existence of metacognition in the metarepresentational sense). I fear that Couchman et al. have engaged in a fine body of experimental work that is framed and guided by theoretical confusion.

R9. Dual processes and judgment

Frankish takes issue with the argument of Section 7 of the target article, which claims that the conscious events that take place at the System 2 level (e.g. verbalizing to myself, "P", or, "I shall do Q") don't have the right kind of causal role to constitute a judgment or a decision. For they only achieve their effects via further (unconscious) processes of reasoning. So although these events are introspectable, this doesn't mean that any judgments or decisions are introspectable. Frankish replies that these events have a *System 2 role* appropriate for a judgment or decision. For they are the *last System 2 events* that occur prior to the characteristic effects of judgments

and decisions. While he acknowledges that further reasoning processes of a System 1 sort occur subsequent to those events, mediating their causal effects on behavior, he says that these should be thought of as belonging to the *realizing base* of a System 2 judgment or decision.

However, our common-sense notions of *judgment* and *decision* don't make any allowance for the System 1 / System 2 distinction. A judgment is a content-bearing event that gives rise to a stored belief with the same content *immediately*, and which is likewise immediately available to inform practical decision-making, without the intervention of any further reasoning. Similarly, a decision is a content-bearing event that causes intention or action without the mediation of any further reasoning about whether or not to act. By these lights, neither the judgment-like event of saying to myself, "P", nor the decision-like event of saying to myself, "I shall do Q", can qualify. Moreover, while it may be true enough that System 2 processes in general are realized in those of System 1 (Carruthers 2009), the realizing conditions for a particular event surely cannot occur subsequent to that event itself. And yet it is only once the conscious events of saying to myself, "P", or, "I shall do Q", are completed that the System 1 reasoning leading to belief or action kicks in. In addition, if we opt to say that the judgment or decision isn't either one of *those* events, but rather the more extended event that also includes the subsequent System 1 practical reasoning, then *that* event isn't an introspectable one. So either way, there is no one event, here, that is both introspectable and is a judgment / decision.

However, let me emphasize that the introspectable events that are involved in System 2 processes are by no means epiphenomenal. On the contrary. Nor, I should stress, is metacognition itself epiphenomenal either, contrary to a claim **Buckner et al.** make about the commitments of the target article. Quite the reverse. System 2 reasoning processes are shot through with—and are largely dependent upon—metacognitive thoughts and beliefs. And on any account, System 2 plays an important part in human cognition and behavior (albeit one that is subject to significant individual differences; see Stanovich 1999).

R10. The evidence from autism

The target article maintains that there is no convincing evidence that in autistic subjects metacognition is preserved while mindreading is damaged (Section 10). This is contrary to the claims of Goldman (2006) and Nichols & Stich (2003), who cite such evidence in support of a "metacognition is prior" account, or a "two independent systems" view, respectively. **Williams et al.** agree with the target article in this respect, and cite Williams & Happé (in press) as demonstrating that autistic children have equivalent difficulty attributing intentions to themselves and to other people, with their performance on these tasks being significantly correlated with their performance on traditional false belief tasks. These new results are very welcome.

However, **Williams et al.** also cite evidence provided by Williams & Happé (under review), which is said to favor a "one system, two modes of access" account over my preferred "mindreading is prior" thesis. In a modified version of the Smarties task autistic children have significantly *greater* difficulty with the "self" version of the task than they do with the "other" version.[4] Williams et al. are mistaken in their interpretation of the significance of their own data, however. This is surprising, since all of the materials for a correct analysis are contained in the very article that they cite (Williams & Happé, under review), as I shall now explain. The upshot is that these new data are fully consistent with a "mindreading is prior" account.

Suppose, first, that autistic children lack the normal mentalizing system altogether. (This seems to be the preferred view of Williams & Happé, under review.) Such children would therefore lack whatever basic "core knowledge", or innate module, or innate domain-specific learning mechanism underlies the development of mentalizing abilities in normal children. Autistic children may nevertheless achieve some limited success in performance by other routes—by means of explicit domain-general theorizing, by memorizing rules and explicit strategies, and so forth. If an account of this sort is correct, then data from autistic subjects are inherently incapable of discriminating between the "mindreading is prior" and the "one mechanism, two modes of access" views of the relationship between mindreading and metacognition. For each of the latter applies only to those people who possess a normal (or near normal) mentalizing *system*, or *faculty*. The "mindreading is prior" account claims that there is just a single mentalizing system, designed initially for mindreading, which is turned upon the self to issue in metacognition. In contrast, the "one mechanism, two modes of access" account, while agreeing that there is just a single mentalizing system, claims that the system in question has both perception-based and introspective channels of access to the mental items in its domain. The former predicts that no one with a normal mentalizing system should possess mindreading competence but lack metacognitive competence; whereas the latter predicts that there might be individuals with a normal mentalizing system who can mindread successfully but who lack a capacity for metacognition, because the introspective channel has been broken or disrupted. Importantly, *neither* model makes any predictions about what might happen in individuals who lack the normal mentalizing system altogether, but who rather "hack" their way to success by other methods. There might be all kinds of reasons why it could be easier to develop rules and strategies that apply to other people than it is to acquire such rules to apply to oneself, as Williams & Happé (under review) themselves argue.

[4] Note that this result is actually the reverse of the claims made by Goldman (2006) and Nichols & Stich (2003). For the data seem to show mindreading relatively intact while metacognition is damaged. **Lombardo et al.** mention similar data in respect of emotion understanding, showing that autistic people do significantly worse on measures of understanding their own emotions than they do on measures of understanding the emotions of others.

Now suppose, in contrast, that autistic children (or at least those who are comparatively high functioning) do possess a mentalizing system, only one that is significantly delayed in its normal development (and is perhaps slower and less reliable in its operations thereafter). And suppose that the "mindreading is prior" account of that system is correct. Still there might be reasons why individuals with a partly formed mentalizing faculty should find some mindreading tasks easier than parallel metacognitive ones. For example, as Williams & Happé (under review) themselves suggest, it may be that the perceptions of action that provide the main input for mentalizing are much more salient and easily accessible in the case of others' actions than in the case of one's own actions.[5]

However, wouldn't such an account predict (contrary to fact) that normally developing children should likewise pass mindreading tasks before they pass the equivalent metacognitive ones? Not necessarily. For the recent data mentioned in R3 suggest that a basic mentalizing *competence* is in place well before children start to be able to pass verbal versions of mentalizing tasks, and that there is some extraneous factor or factors that inhibit verbal performance. And the latter might contain no bias in favor of "other" versus "self" versions of the tasks. In the case of autistic children, in contrast, it is the delayed development of the mentalizing system itself that delays successful performance, enabling a bias in favor of mindreading over metacognition to display itself.

R11. Neuroimaging evidence

Lombardo et al. cite neuroimaging evidence showing that identical neural regions are implicated in mentalizing about self and other, and that there are no other areas of the brain that are recruited specifically for mentalizing about self, or about other. These data are very welcome, and provide strong support for the "mindreading is prior" model. This is because all three of the competing accounts predict that there should be some brain regions used specifically for mentalizing about oneself and/or brain regions used specifically for mentalizing about others. Lombardo et al. claim, however, that it is an implication of the "mindreading is prior" account that the various brain regions implicated in mentalizing should be activated *to the same degree* when mentalizing about the self or about another. Since their data conflict with this prediction, they take this to raise a puzzle for the view. However, the "mindreading is prior" account makes no such prediction. For it allows that different kinds of data are implicated in the two forms of mentalizing. Specifically, mentalizing about the self can utilize visual and auditory imagery, somatosensory experiences, and so forth, in a way that mentalizing about others normally can't. I

[5] Note, too, that explanations similar to those provided here can accommodate the data cited by **Lombardo et al.** on autistic people's differential understanding of emotion in self and other.

suggest that these differences are sufficient to explain the different degrees of neural activation in question.

Nor, it should be stressed, does the "mindreading is prior" account predict that mindreading tasks are always performed in the same way (on the contrary, see R2). So the findings reported by **Lombardo et al.**—that people tend to rely more on stereotypes when reasoning about the mental states of dissimilar others, while utilizing simulation strategies when reasoning about the mental states of people who are perceived to be similar to themselves—raise no particular challenge for the account.

R12. The evidence from schizophrenia

The target article discusses the significance of the finding that schizophrenic patients with "passivity" symptoms have difficulties in attributing intentions to themselves while being normal at reading the minds of others. Nichols & Stich (2003) argue that this reveals a dissociation between metacognitive and mindreading abilities, whereas the target article suggests that the data are better explained in terms of *faulty* or *unusual* experiences being presented as input to an intact mindreading system. In contrast, **Wiffen & David** cast doubt upon the reliability of the data in question. If they are right, then that makes a "mindreading is prior" account even easier to defend.

Robbins, on the other hand, argues that schizophrenic patients with paranoid symptoms seem to display the contrary dissociation. For such patients perform poorly in mindreading tasks of various sorts, whereas there is no evidence (he tells us) that they show equivalent metacognitive deficits. **Wiffen & David** present two such strands of evidence, however. One is that schizophrenic patients characteristically lack insight into their own condition, which is (Wiffen & David claim) a failure of metacognition. But here they weave a tangled story. For while most if not all schizophrenic patients *do* perform poorly on tests of mindreading and *do* lack insight into their own illness, they appear to have no difficulties in distinguishing between normal and psychotic thoughts, feelings, and behavior in another person (Startup 1997). This raises a puzzle. If the lack of insight that these patients have into their own condition results from poor mindreading abilities, then how is it that they can nevertheless possess insight into the disordered minds of others?

We can begin to unravel this puzzle by noting that even if paranoid beliefs result partly from faulty mindreading, they cannot result from faulty mindreading *alone*. There must also exist a willingness to believe propositions whose prior probability is very low, in some circumstances. (Most of us may have entertained a paranoid *thought* or *hypothesis* at one time or another, but

have immediately dismissed the idea as absurd.) And indeed, there is an extensive body of literature demonstrating that people with schizophrenia display a marked "jumping to conclusions" bias, forming beliefs from new data much more swiftly and with higher degrees of confidence than do controls. (See Blackwood et al. 2001 for a review.) Moreover, the bias in question seems to be one of data-gathering rather than a failure of probabilistic reasoning as such, since patients with schizophrenia reason normally about the plausibility of hypotheses that are presented to them, or when presented with the same range of data that lead normal individuals to formulate a new belief. This could explain why patients with schizophrenia lack insight into their own condition while showing insight into the conditions of others. For in the first case they are *forming* a paranoid belief from limited data, whereas in the latter case they are assessing the prior probability of someone else's belief.

Wiffen & David's other strand of evidence suggesting that schizophrenic patients have parallel difficulties in mindreading and metacognition is much more direct. They cite the demonstration by Koren et al. (2004) that schizophrenic patients do poorly on tests of metacognitive ability. (See also Koren et al. 2006.) Specifically, Koren et al. administered the Wisconsin Card Sorting Test to patients with schizophrenia, while also asking them to provide confidence ratings of their recent choices and while allowing them to decide whether or not each sorting of the cards should count towards their final score (and potential monetary gain). The patients performed normally on the test itself, but displayed a marked deficit on the metacognitive measures. While this isn't yet an extensive body of data, it does suggest that deficits in mindreading and metacognition are paired together in schizophrenia, just as the "mindreading is prior" account would predict.

R13. Some friendly suggestions

A number of commentaries are entirely friendly to the approach taken in the target article, and hence need only a brief mention. First, **Evans** uses the position defended in the target article to resolve a tension in many theorists' thinking about dual systems of reasoning. For System 2 is often characterized as a conscious system, whereas we know that people's reports of System 2 processes are often confabulated. The solution is to note that only the globally broadcast contents of working memory are ever accessible to the mindreading system that is responsible for self-report, whereas many other aspects of System 2 processing will remain inaccessible to it. The contents of working memory represent but small islands of consciousness within the overall operations of System 2, leaving plenty of scope for confabulation about the remainder.

Second, **Huebner & Dennett** emphasize the dangers inherent in the use that is made of first-person pronouns throughout the target article, as in, "I have access to my own visual images", or, "We do have introspective access to inner speech." For these seem to imply a place for the *self* in

the account, in addition to the various subpersonal systems described (for language, for mindreading, and so forth). Of course I intend no such thing. The outputs of the mindreading system are passed along as input to a variety of other systems, included in which is a language production mechanism that might issue in a (covert or overt) expression of the metarepresentational content in question; that is all. While use of personal pronouns in cognitive science is a handy *façon de parler*, we need to take care that their use is eliminable from the theories in question. I have no doubt, however, that they can be eliminated from all aspects of the "mindreading is prior" account.

Third, **Langland-Hassan** offers a welcome corrective to what I actually wrote in the target article, though not to anything that I believe or really intended to say. I had claimed that perceptual and quasi-perceptual states can be self-ascribed without interpretation by virtue of being globally broadcast. But Langland-Hassan points out that the question whether the speech that I seem to hear running through my head is my own or is really the voice of another person cannot be answered without interpretation. For by hypothesis the mindreading system has no access to my own articulatory intentions. All it has access to is the resulting experience. Likewise for the question whether a visual image that I am currently entertaining is a memory-image or a fantasy-image. No experience can wear its own provenance on its face. Hence describing myself as *remembering* the event depicted will have to be based on an inference grounded in aspects of the immediate context, feelings of familiarity, and so forth. All of this is entirely correct. What I should have said is that the *contents* of globally broadcast states can be self-attributed without interpretation, but interpretation is required for one to know to what *kind* those states belong. This leaves untouched the claim that the mindreading system has accessible to it data that it can use when self-ascribing propositional attitude states that is of no help in ascribing such states to other people.

R14. Behaviorism bites back

Catania offers a behaviorist alternative to my account, citing the work of Skinner (1945, 1963). Likewise, **Lewis & Carpendale** challenge the computationalist assumptions made by the target article, while criticizing me for not taking account of the work of the later Wittgenstein. I don't believe that I should need to argue in support of either cognitivism or computationalism, since both are foundational assumptions of most of cognitive science. In any case I don't have the space to defend them here. (See Gallistel & King 2009 for the definitive argument.) In addition, I don't believe that Wittgenstein's work contains any challenges that cognitive science cannot easily answer. There is some irony, moreover, in the charge that I should have paid more attention to Wittgenstein. For I spent the first fifteen years of my academic career focused on his philosophy, and much of that time was devoted to the so-called "private language argument" that

Lewis & Carpendale admiringly refer to. This formed the topic of my doctoral dissertation. I ultimately came to believe that no version of the argument can be successful that doesn't already rely on anti-realist (e.g. behaviorist) or verificationist premises.

R15. Conclusion

I am grateful to my commentators for the care and attention that they devoted to the target article. As a result, the theoretical options have been further clarified, and the "mindreading is prior" model of self-awareness has been additionally elaborated and strengthened. At the very least, that model will now need to be taken seriously by anyone considering the nature of self-awareness and its relationship to our mindreading abilities. And now that the strengths and weaknesses of the four main theoretical options have been clearly laid out, there is an urgent need for additional experimental data that will enable us to discriminate between them. As things stand, my own verdict is that the "mindreading is prior" account is the one that is best supported by the existing evidence (in part because it is the most parsimonious). But future findings could change all that.

References

Note: the list below only includes items not already listed in the original target article or in one or more of the commentaries.

Apperly, I., Riggs, K., Simpson, A., Chiavarino, C., & Samson, D. (2007). Is belief reasoning automatic? *Psychological Science*, 17, 841-844.

Baddeley, A. (1990). *Human Memory*. Lawrence Erlbaum.

Baron-Cohen, S. (1995). *Mindblindness*. MIT Press.

Block, N. (1986). An advertisement for a semantics for psychology. In P. French, T. Euhling, & H. Wettstein (eds.), *Midwest Studies in Philosophy*, vol. X, *Studies in the Philosophy of Mind*, University of Minnesota Press.

Bloom, P. (2002). *How Children Learn the Meaning of Words*. MIT Press.

Botterill, G. & Carruthers, P. (1999). *The Philosophy of Psychology*. Cambridge University Press.

Csibra, G., Bı´ro´, S., Koo´s, O., & Gergely, G. (2003). One-year-old infants use teleological representations of actions productively. *Cognitive Science*, 27, 111–133.

Dunlosky, J. & Metcalfe, J. (2009). *Metacognition.* Sage.

Flavell, J. (1979). Metacognition and cognitive monitoring: A new era of cognitive-developmental inquiry. *American Psychologist*, 34, 906-911.

Fodor, J. (1990). *A Theory of Content and Other Essays*. MIT Press.

Fodor, J. (1992). A theory of the child's theory of mind. *Cognition*, 44, 283-296.

Gallistel, C.R. & King, A. (2009). *Learning, Memory, and the Brain*. Wiley–Blackwell.

Johnson, S. (2000). The recognition of mentalistic agents in infancy. *Trends in Cognitive Sciences*, 4, 22–28.

Keysar, B., Lin, S., & Barr, D. (2003). Limits on theory of mind use in adults. *Cognition*, 89, 25-41.

Koren, D., Seidman, L., Goldsmith, M., & Harvey, P. (2006). Real world cognitive—and metacognitive—dysfunction in schizophrenia: a new approach for measuring (and remediating) more "right stuff". *Schizophrenia Bulletin*, 32, 310-326.

Koriat, A. (2007). Metacognition and consciousness. In P. Zelazo, M. Moscovitch, & E. Thompson (eds.), *The Cambridge handbook of consciousness*, Cambridge University Press.

Luo, Y. & Baillargeon, R. (2005). Can a self-propelled box have a goal? Psychological reasoning in 5 month old infants. *Psychological Science*, 16, 601-608.

Metcalfe, J. (2008). Evolution of metacognition. In J. Dunlosky & R. Bjork (eds.), *Handbook of Metacognition and Learning*, Lawrence Erlbaum.

Scholl, B. (2007). Object persistence in philosophy and psychology. *Mind and Language*, *22*, 563-591.

Siegal, M. & Varley, R. (2002). Neural systems involved in theory of mind. *Nature Reviews Neuroscience*, 3, 462-471.

Varley, R. (1998). Aphasic language, aphasic thought. In P. Carruthers & J. Boucher (eds.), *Language and Thought*, Cambridge University Press.

Woodward, A. (1998). Infants selectively encode the goal object of an actor's reach. *Cognition*, 69, 1-34.